Temporal predictive regression models for linguistic style analysis

Carmen Klaussner and Carl Vogel
School of Computer Science and Statistics,
Trinity College Dublin

ABSTRACT

This study focuses on modelling general and individual language change over several decades. A timeline prediction task was used to identify interesting temporal features. Our previous work achieved high accuracy in predicting publication year, using lexical features marked for syntactic context. In this study, we use four feature types (character, word stem, part-of-speech, and word n-grams) to predict publication year, and then use associated models to determine constant and changing features in individual and general language use. We do this for two corpora, one containing texts by two different authors, published over a fifty-year period, and a reference corpus containing a variety of text types, representing general language style over time, for the same temporal span as the two authors. Our linear regression models achieve good accuracy with the two-author data set, and very good results with the reference corpus, bringing to light interesting features of language change.

Keywords:
language change,
style analysis,
regression

1 INTRODUCTION

Statistical style analysis or 'stylometry' is the automatic analysis of authorial style, usually investigating the frequency of occurrence of specific features in a given author's works. Features with consistent frequencies are assumed to be representative of that author, and features are also considered discriminative if other comparable authors use them with consistently different frequencies. This type of analy-

sis is known as synchronic analysis, as it disregards composition or publication dates.

However, this is a simplification, since most writers compose over time spans of 20–40 years, where they not only undergo individual stylistic development, but also bear witness to general contemporaneous language change. These two types of temporal influences can cause synchronic analyses to be misinterpreted. Thus, as already discussed by Daelemans (2013), unless style is found to be invariant for an author and does not change with age and experience, temporality can be a confounding factor in stylometry and authorship attribution. For this reason, diachrony presents an important aspect of style analysis, not only to disambiguate synchronic analyses of style, but also in its own right by modelling language change over time.

In this work, we examine language change in two literary authors, as well as the corresponding background language change during the same time period. Specifically, we are interested in features that are attested in each time slice of the diachronic corpus studied. We refer to this subset of features that appear in all samples as 'constant' features. This classification captures occurrence patterns rather than variation in terms of relative frequencies, which may or may not change over the time intervals examined. In order to identify salient constant features that exhibit change over time, we refer to a temporal prediction task based on the features' relative frequencies.

This extends our previous work on predicting the publication year of a text using syntactic word features (Klaussner and Vogel 2015).[1] That study considered a data set comprising works by two authors from the 19th to the 20th century, as well as a data set based on a reference corpus, and sampled features that appeared in many, but not necessarily all, time slices. For the two-author data set, a root-mean-square error (RMSE) of 7.2 years[2] on unseen data (baseline: 13.2) was

[1] These are lexical features that have been marked for syntactic function to differentiate between lexical representations that can appear in different syntactic contexts (see Section 4.2).

[2] Hereafter, when we report RMSE, we take the units to be years and do not repeat the unit. This is to be understood with the caveat that the data are processed using only integer values of years. Temporal prediction for any text cannot be wrong by '7.2 years', but rather by seven or eight years. The RMSE is an aggregate.

obtained, whereas the model built on the larger reference data set obtained an RMSE of 4 on unseen data (baseline: 17). While the current work is similar in that it uses the same data sets and the same general prediction task, it is different in that achieving 'high accuracy' of prediction is not the main objective here. Although we report our results and compare them to those from the earlier study, the prediction task is primarily used as a means to determine what is stable and what changes in individual and general language use over time.[3] Hence, the purpose is not the pursuit of a perfect temporal classifier, but rather to understand 'typical' distributions of linguistic feature categories during an author's lifetime. This change must also be understood in relation to the effects of ageing on language production, as explored for instance by Pennebaker and Stone (2003). Features that are not constant in the sense analysed here are also important. We focus on constant features, because if they are used in each time slice throughout an author's career, then they are probably integral to that author's style, making the relative frequencies of such features across time slices interesting to explore.

The contribution of this new study is the analysis of language change using an extended feature set, adding character,[4] word stem, and syntactic (part-of-speech tag) features to the previous set, which consisted only of syntactic word features. In addition, rather than considering only unigram size, this study analyses all n-gram sizes up to length four. Therefore, one of the questions investigated as part of this work is whether (and to what extent) the more linguistically informative features, such as syntactic word n-grams, exhibit more dramatic change than lexicographic and part-of-speech features. We present our own method for reasoning about temporal change in constant linguistic features, using standard techniques from regression analysis, particularly parameter shrinkage.[5] We find that the best predictive values common to the works by the two authors and the reference corpus are word stem, and POS bigrams and trigrams, which also account for

[3] The data sets for the two authors are analysed both separately and together.

[4] This feature type covers alphanumeric characters, punctuation, and spaces.

[5] The resulting set of features identified is a specific subset of features that are both constant and have a linear relationship with the response variable over time, i.e. a change in trend rather than in periodicity. Non-linear patterns or estimation may also be interesting, but our focus is different here.

most shared model predictors. In terms of language change, with the help of our regression models, we identified several differences between the reference corpus and the works by the two authors.

The remainder of this article is structured as follows: Section 2 outlines previous work in the area; Section 3 discusses methods; Section 4 presents the data sets, preprocessing steps, and feature types; Section 5 discusses the general experimental setup and the experiments themselves. Section 6 reports and analyses the salient features of the models. Section 7 discusses the results, and Section 8 concludes this work.

2 RELATED WORK

Studies in the field of style analysis or 'stylometry' focus on different sub-tasks, such as authorship attribution; i.e. given an unknown document and several candidate authors, the task is to decide which candidate is most likely to have authored the document. This problem can be studied in a closed-class or open-class scenario. The former assumes that the true author is among the set of candidates, rendering the task of determining who authored the document in question simpler than in the open-class variant, where the set of candidates may or may not contain the true author. Open-class authorship attribution has been studied for instance by Koppel *et al.* (2011), who consider authorship attribution in the presence of what they conceive are the three most common deterrents to using common authorship techniques, i.e. possibly thousands of known candidate authors, the author of the anonymous text not being among the candidates, and the 'known-text' for each candidate and/or the anonymous text being very limited. Considering a set of blog posts (extracting 2,000 words for the known text and a 500-word-long test snippet), they use a similarity-based approach (cosine similarity) on space-free character tetragrams. The task is to find the author of a given text snippet, based on evidence from varying feature sets, the rationale being that only the right author is going to be consistently similar to his or her own 'unknown' piece. An author is selected only if above a particular proportion or threshold, otherwise the method returns a 'Don't know' answer. Unsurprisingly, a greater number of feature sets and a closed-candidate set yield greater accuracy, i.e. 87.9% precision with 28.2% recall. In

the closed-candidate setting, reducing the number of candidates improves accuracy (e.g. 1,000 candidates yields 93.2% precision with 39.3% recall), whereas in the open-class setting, having fewer candidates actually introduces problems, in that an author might end up being chosen erroneously, because there is less competition. Overall, Koppel *et al.* (2011) find that their methods achieve passable results even for snippets as short as 100 words, but note that there is still no satisfactory solution for the case of a small open-candidate set and limited anonymous text.

Another general variant of the attribution problem is commonly referred to as 'Authorship verification', which requires determining whether a piece of text has been written by a specific author. This has been considered by Koppel *et al.* (2007), for instance, who show that the task of deciding whether an author has written a particular text can be accurately determined by iteratively removing the set of best features from the learning process: the differences between two texts by the same author are usually only reflected in a relatively small number of features, causing accuracy to drop much faster and more dramatically than when the texts were not written by the same person. In contrast, 'Author profiling', which involves predicting an author's characteristics, such as gender, age or personality traits, based on a particular text, has been studied extensively as part of the PAN competitions (e.g. see Rosso *et al.* 2016). While the predicted variable varies by task, what is common to the studies above as well as to our own is the use of relative frequencies of some feature to predict the variable of interest, using similarity-based or statistical methods.

However, while the general scenario is the same, diachronic studies differ in that they take into account the temporal ordering of an author's works, seeking to reveal temporal changes within his or her style rather than changes between authors or between different texts by the same author. A few works focus more specifically on temporality in style analyses. Previous work by Smith and Kelly (2002) investigates the question of whether vocabulary richness remains constant over time, by examining measures of lexical richness across the diachronic corpora of three playwrights (Euripides, Aristophanes, and Terence). The plays are divided into standardized non-overlapping blocks, each being analysed for certain properties pertaining to lexical richness, such as vocabulary richness, pro-

portion of *hapax legomena*, and repetition of frequently appearing vocabulary. In addition to testing the constancy of these properties over time, weighted linear regression is used to test associations between these measures and the time of a play's first performance. For this, the property's value in a particular text block is used as response, and time of performance is used as predictor.[6] Results show that Aristophanes' use of *hapax legomena* appears to have decreased over time. Interestingly, one of his earlier works, *Clouds*, which was subjected to redrafting after the first staging, but for which the finishing date is unknown, is predicted to originate towards the end of the playwright's life, indicating that revisions might have been made at a much later stage. Our work here also uses linear regression, but rather than using time as predictor, we investigate to what extent pooled information from several features can accurately predict a text's publication year. The study presented by Hoover (2007) considers language change in Henry James' style with respect to the 100–4,000 most frequent word unigrams, using methods such as 'Cluster Analysis', 'Burrows' Delta', 'Principal Component Analysis', and 'Distinctiveness Ratio'.[7] Three different divisions, into early (1877–1881), intermediate (1886–1890), and late style (1897–1917), emerge from the analysis.[8] However, rather than being strict divisions, there seem to be gradual transitions, with the first novels of the late period being somewhat different from the others, suggesting that it might be interesting to conduct a continuous analysis of style in James' works. Thus, in contrast to the previous study, the work we present here focuses on a more graduated interpretation of style over time, with yearly intervals rather than classification into

[6] In order to perform inverse prediction, i.e. predicting the date of an unknown work by the measure, the authors draw a horizontal line at y, with y corresponding to the measure's average in the text and look at the intersection with the estimated regression line.

[7] Distinctiveness Ratio: Measure of variability defined by the rate of occurrence of a word in a text divided by its rate of occurrence in another. Principal Component Analysis (PCA) is an unsupervised statistical technique to convert a set of possibly related variables to a new uncorrelated representation, i.e., principal components.

[8] The same divisions have also been identified by literary scholars (Beach 1918).

different periods along the timeline of the author's works. Our work on temporal prediction (Klaussner and Vogel 2015) considered the task of accurately predicting the publication year of a text through the relative frequencies of syntactic word features.[9] We used multiple linear regression models to predict the year when a text was published, for three data sets, the first containing texts by Mark Twain and Henry James, the second a mid 19th to early 20th century reference corpus, and a third one combining all data from the previous two sets. Although the data for the two authors had been kept separate to allow for potentially different levels between them, the models disregarding authorial source tended to be more accurate (RMSE of 7.2 vs. 8.0). While the reference corpus model performed well on its own test set (RMSE of 4), using it to predict publication year for the two authors was rather inaccurate (RMSE: 15.4 for Twain, and 20.3 for James). This suggests that the style of the two authors was rather different from general language, Twain's being somewhat more similar to it than James'. Combining all data leads to more accurate results (RMSE: 1.8), and model features and estimates suggest a marked influence of Twain and James on the model, in spite of their smaller data sets (for more detailed, quantitative results, see Section 5.3).

On the topic of suitable stylistic feature types in this context, Stamatatos (2012) compares the performances of the most frequent function words and character trigrams for the authorship attribution task. It is shown that character trigrams outperform word features, especially when training and test corpus differ in genre – they are also found to be more robust and effective when considering different feature input sizes. For this reason, we include character n-grams as a feature type here as well. In contrast to part-of-speech tags or word stems, character n-grams present a less linguistically motivated feature type, as writers would not be able to control the number of times a particular character is used to the same extent as they would be able to control their choice of particular syntactic constructions. Yet this feature type becomes more likely to bear meaning, as character n-gram size increases, approaching average word length.

[9] Syntactic word features are words marked for their syntactic context. This is explained in more detail in Section 4.2.

3 METHODS

This section discusses the methods used in this work, beginning with temporal regression models (Section 3.1), and continuing with evaluation techniques for these predictive models (Section 3.2).

3.1 *Temporal regression models*

The analysis of data over time probably has its most prominent usage in quantitative forecasting analysis, which involves the (quantitative) analysis of how a particular variable (or variables) may change over time and how that information can be used to predict its (or their) future behaviour, thus inherently assuming that some aspects of the past continue in the future, known as the 'continuity assumption' (Makridakis *et al.* 2008). Thus, a future value of a variable y is predicted using a function over some other variable values. These other variable values could be composed in two different ways, pertaining either to the use of a 'time-series' model or an 'explanatory' model. When considering a time-series model, the assumption is that one can predict the future value of the variable y by looking at the values it took at previous points in time and the possible patterns this would show over time. In contrast, for prediction, explanatory models focus less on interpreting previous values of the same variable, and more on the relationship with other variables at the same point in time. Consequently, the prediction of a variable y, using explanatory models, is based on a function over a set of distinct variables: $x_1, x_2, \ldots, x_{p-1}, x_p = X$, with $y \notin X$, at the same time point $t : \{t \in 1, \ldots, n\}$, and some error term: $y_t = f(x_{1t}, \ldots, x_{2t}, \ldots, x_{p-1t}, \ldots, x_{pt}, error)$.

The general model for this is shown in Equation (1), predicting variable y, where $\hat{y}_t$ refers to the estimate of that variable at a particular time instance $t : \{t \in 1, \ldots, n\}$, β_0 refers to the intercept, and β_p to the pth coefficient of the pth predictor x_{pt}.

$$\hat{y}_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} \tag{1}$$

In the present case, the year of publication is always set as the response variable, so that a model based on syntactic unigrams (relative frequencies) for the year 1880 could be defined in the following way: $\hat{y}_{1880} = \beta_0 + \beta_1(NN_{1880}) + \beta_2(NP_{1880}) + \beta_3(IN_{1880})$.

Regression models are customarily evaluated using the residual sum of squares (RSS): given predicted values $\hat{y}_i$ computed by the

model and observed values y_i, the RSS measures the difference between them. The smaller the RSS, the greater the amount of variation of y values around their mean that is explained by the model. This is known as the 'ordinary least squares' (OLS) fit, a model selection criterion that also forms the basis of evaluation measures, such as the root-mean-square error (RMSE) (see Section 3.2).

In this work, rather than applying models based only on least squares regression, we employ so-called 'shrinkage' models that offer an extension to regular OLS models by additionally penalizing coefficient magnitudes, thus aiming to keep the model from overfitting the data. Specifically, we use the 'elastic net', which is a combination of the two most common types of shrinkage, 'lasso' and 'ridge' regression (Zou and Hastie 2005). The elastic net penalizes both the L_1 and L_2 norms,[10] causing some coefficients to be shrunk (ridge) and some to be set to zero (lasso), with the exact weighting between the two also being subject to tuning. In addition, the elastic net tends to select groups of correlated predictors rather than discarding all but one from a group of related predictors, as is common when using only the lasso technique. The entire cost function is shown in Equation (2). As with the lasso and ridge regression, $\lambda \geq 0$ controls finding a compromise between fitting the data and keeping coefficient values as small as possible, while the elastic net parameter α determines the mix of the two penalties, i.e. how many features are merely shrunk as opposed to being completely removed.

$$\max_{\{\beta_{0,k},\beta_k \in \mathbf{R}^p\}_1^K} \left[\sum_{i=1}^{N} \log \Pr(g_i|x_i) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} (\alpha|\beta_{kj}| + (1-\alpha)\beta_{kj}^2) \right] \tag{2}$$

There are numerous advantages to using shrinkage models, and the elastic net estimation in particular, such as built-in feature selection and more robust and reliable coefficient estimation. This is discussed in more detail for instance by James *et al.* (2013, pp. 203–204) and Friedman *et al.* (2001, pp. 662–663).

3.2 *Evaluation*

The 'root-mean-square error' (RMSE) is one of the measures that can be used for the purpose of evaluating linear regression models: it is

[10] $\|\beta\|_1$: $\sum_i |\beta_i|$ and $\|\beta\|_2^2$: $\sum_i \beta_i^2$

defined as the square root of the variance of the residuals between outcome and predicted value and thus provides the standard deviation around the predicted value, as shown in Equation 3.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{3}$$

The advantage over the more general 'mean-square error' (MSE) is that RMSE computes deviations in predictions on the same scale as the data. However, due to the squaring, assigning more weight to larger errors, the RMSE is more sensitive to outliers.

4 DATA

The following section presents the data sets (Section 4.1), followed by feature types (Section 4.2), and finally, data preparation (Section 4.3).

4.1 *Data sets*

The data for this study originates from three separate sources: works by two American authors, Mark Twain and Henry James, and a reference corpus for American English, from 1860 to 1919.

Mark Twain and Henry James were chosen for this analysis because both were prolific authors writing over a similar time span, from the late 19th to the early 20th century. The study presented by Hoover (2007), mentioned in Section 2, provided the first evidence that a temporal analysis of James' work might be fruitful; other sources (Beach 1918; Canby 1951) indicated that it might be interesting to study works by Henry James and Mark Twain, two highly articulate and creative writers, contrasting in temperament and in their art (Canby 1951, p. xii), yet each conscious of the other (Brooks 1920; Ayres 2010). Considering individual authors might be more interesting from an interpretative viewpoint, in that the phenomena observed are more likely to be directly attributable to the author(s) examined. However, one needs a reference corpus representing 'average' style to know what importance to assign to a particular phenomenon. For instance, one might discover a decrease in usage of a particular feature over time for Twain and James; if the same feature also decreased in usage in general, this discovery would not necessarily be noteworthy. While

both individual and general language change are of interest in their own right, they also provide comparative information about the relative importance of the features observed, indicating whether particular events are likely to be unusual.

For each of the two authors, we compiled a separate data set of their main works.[11] Table 1 shows the data for Henry James, and Table 2 that for Mark Twain. The texts were collected from the *Project Gutenberg*[12] and the *Internet Archive*[13] selecting the earliest editions available. The reference corpus was assembled by taking an extract from *The Corpus of Historical American English* (COHA; Davies 2012).[14] The COHA is a 400-million-word corpus, containing samples of American English from 1810–2009, balanced in size, genre and sub-genre in each decade (1,000–2,500 files each). It contains balanced language samples from fiction, popular magazines, newspapers and non-fiction books, which are again balanced across sub-genres, such as drama and poetry.[15] The COHA data were compiled from different sources, some of which were already available as part of existing text archives (e.g., *Project Gutenberg* and *Making of America*), whereas others were converted from PDF images, or scanned from printed sources. The corpus allows analysis of linguistic change at different levels, i.e. lexical, morphological, syntactic, and semantic.

4.2 *Feature types*

For the experiments described in Section 5, we consider four different types of features, as well as various sequence sizes of these. Table 3 lists all feature types, ordered by increasing degree of specificity, with an example for unigrams, and one for trigrams.

The most general type is character n-grams, including punctuation and single spaces.[16] While the character n-grams reduce words

[11] In this case, 'main' is with reference to the size of the work in kilobytes, rather than in terms of literary importance. We use kilobytes instead of word count, as this gives a more precise indication of file size.

[12] `http://www.gutenberg.org/` – last verified March 2018.

[13] `https://archive.org/` – last verified March 2018.

[14] Free version available from: `http://corpus.byu.edu/coha/` – last verified March 2018.

[15] An Excel file with a detailed list of sources is available from: `http://corpus.byu.edu/coha/` – last verified March 2018.

[16] Multiple spaces were reduced to single spaces.

Table 1: Collected works for Henry James. Showing 'Title', the original publication date ('1st Pub.'), version collected ('Version'), 'Size' in kilobytes and 'Genre' type. The dashed lines indicate the boundaries for compression, i.e. which of the works are combined into one temporal interval (see Section 4.3 for discussion of the compression technique used)

Title	1st Pub.	Version	Size	Genre
The American	1877	1877	721	novel
Watch and Ward	1871	1878	345	novel
Daisy Miller	1879	1879	119	novella
The Europeans	1878	1879	346	novel
Hawthorne	1879	1879	314	biography
Confidence	1879	1880	429	novel
Washington Square	1880	1881	360	novel
Portrait of a Lady	1881	1882	1200	novel
Roderick Hudson	1875	1883	750	novel
The Bostonians	1886	1886	906	novel
Princess Casamassima	1886	1886	1100	novel
The Reverberator	1888	1888	297	novel
The Aspern Papers	1888	1888	202	novella
The Tragic Muse	1890	1890	1100	novel
Picture and Text	1893	1893	182	essays
The Other House	1896	1896	406	novel
What Maisie Knew	1897	1897	540	novel
The Spoils of Poynton	1897	1897	376	novel
In the Cage	1893	1898	191	novella
Turn of the Screw	1898	1898	223	novella
The Awkward Age	1899	1899	749	novel
Little Tour in France	1884	1900	418	travel writings
The Sacred Fount	1901	1901	407	novel
The Wings of the Dove	1902	1902	1003.7	novel
The Golden Bowl	1904	1904	1100	novel
Views and Reviews	1908	1908	279	literary criticism
Italian Hours	1909	1909	711	travel essays
The Ambassadors	1903	1909	890	novel
The Outcry	1911	1911	304	novel
*The Ivory Tower**	1917	1917	488	novel
*The Sense of the Past**	1917	1917	491	novel

'*' indicates unfinished works.

Table 2: Collected works for Mark Twain. Showing 'Title', the original publication date ('1st Pub.'), version collected ('Version'), 'Size' in kilobytes and 'Genre' type. The dashed lines indicate the boundaries for compression, i.e. which of the works are combined into one temporal interval (see Section 4.3 for discussion of the compression technique used)

Title	1st Pub.	Version	Size	Genre
Innocents Abroad	1869	1869	1100	travel novel
The Gilded Age: A Tale of Today	1873	1873	866	novel
The Adventures of Tom Sawyer	1876	1884	378	novel
A Tramp Abroad	1880	1880	849	travel literature
Roughing It	1880	1880	923	semi-autobiog.
The Prince and the Pauper	1881	1882	394	novel
Life on the Mississippi	1883	1883	777	memoir
The Adventures of Huckleberry Finn	1884	1885	586	novel
A Connecticut Yankee in King Arthur's Court	1889	1889	628	novel
The American Claimant	1892	1892	354	novel
The Tragedy of Pudd'nhead Wilson	1894	1894	286	novel
Tom Sawyer Detective	1896	1896	116	novel
Personal Recollections of Joan Arc	1896	1896	796	historical novel
Following the Equator	1897	1897	1000	travel novel
Those Extraordinary Twins	1894	1899	120	short story
A Double Barrelled Detective Story	1902	1902	103	short story
Christian Science	1907	1907	338	essays
Chapters from My Autobiography	1907	1907	593	autobiog.
*The Mysterious Stranger**	1908	1897–1908	192	novel

'*' indicates unfinished works.

and sentences to their orthography, the part-of-speech (POS) type generalizes them as sequences of syntactic types. Word stems present a more specific generalization of the simple word feature, but rather than capturing syntactic aspects, this type captures what lexical type of word (or sequence) was used, such as ⟨allud to⟩ in place of 'allude

Table 3: Feature types

n-gram type	Example	
	unigram	*trigram*
character	⟨c⟩	⟨ca,⟩
part-of-speech (POS)	⟨NP⟩	⟨IN DET NP⟩
word stem	⟨allud⟩	⟨to allud to⟩
syntactic word (lexical)	⟨like.IN⟩	⟨like.VB the.DET others.NNS⟩

to' or 'alludes to'.[17] The most specific is termed 'syntactic word' sequences, meaning words that have been marked for syntactic class, as in the case of 'like', which may be used as a preposition or a verb, depending on context. Compare *I'm like my father.* and *I like my father.*: in the first instance 'like' is used as a preposition, in the second it is used as a verb. Hence, for this feature type, each word is given the correct part-of-speech tag, thus allowing distinct features to be identified for words with more than one syntactic context, such as ⟨like.VB⟩ for verbal usage and ⟨like.IN⟩ for prepositional usage.

4.3 *Data preparation*

Before features could be extracted from the two authors' texts, each file had to be checked manually, to remove parts that were written at a different time from the main work, or introductions or comments not by the author, such as notes or introductions by editors. Following this, all source files were then searched (both automatically and manually) to remove unwanted formatting sequences and to normalize spacing.[18]

To extract both POS and syntactic word features, we used the TreeTagger POS tagger (Michalke 2014; Schmid 1994). The original word plus its tag is retained for syntactic word features, while for POS features, the original word is replaced by the POS tag.[19] After ex-

[17] The feature remains orthographic inasmuch as the stem differs from the lemma.

[18] The package *stylo* (Eder *et al.* 2013) was used to convert words into character sequences, while the *RTextTools* package (Jurka *et al.* 2012) was used to extract word stems.

[19] Punctuation and sentence endings are also included as features and in relativization. The POS tags assigned by the tagger to the individual word entity in its context are used to augment or replace the word entity. Individual entities within ⟨...⟩ are separated by a space.

traction, all feature types were then transformed to lowercase, as for this work we do not analyse features with respect to sentence boundaries. Finally, document-feature matrices were constructed for each type and n-gram size and relativized in the following way: for all of the analyses reported here, we compute relative frequencies to take into account any differences in the amount of text available for each year.[20] If more than one work was available for a particular year and authorial source, they were joined together and relativized as one text. For both the reference set and the two-author set, an ordinal variable 'year' was added for each experiment to mark the publication year of a text. The data sets for the two authors were joined into one set after relativization, with an additional categorical variable 'author' to mark which author composed the text. In some instances, both authors published work during the same year; the 'author' variable served to keep such cases separate. Thus, detecting differences in levels of relative frequency by author remains possible within the joint data set. Combined relativization might distort individual interpretation or create a shift towards the author with more data in a given year. The model is trained on 'combined' data, in the sense that there may be two relative frequencies contributing observations to one predictor variable. The categorical author variable may be added to the model, if the level for that predictor differs between James and Twain.

5 EXPERIMENTS

Section 5.1 addresses general experimental design, and model and parameter selection. The four feature types described in Table 3 are considered separately for the two data sets hereafter, with Section 5.2 presenting the results, and Section 5.3 comparing them with the previous study.

5.1 *Model computations*

Before the experiments, the same procedure was performed for all of the previously constructed document-feature matrices, to construct

[20] Long and rarer n-gram sequences could cause the data to become rather sparse and feature values could thus become computationally expensive. To overcome this challenge, memory-intensive processing steps were separated and simplified, using the R packages *bigmemory* (Kane *et al.* 2013) and *foreach* (Revolution Analytics and Weston 2014).

the input for each of the 32 models shown in Table 5. The data were first divided into training and test data using a 75/25 stratified split on the ordinal variable 'year' that we added at the previous step.[21] After that step, we extracted all constant features from the training set, i.e. the features appearing in all training set instances, which were then passed to the elastic net models.[22]

The final model was then computed by performing 10-fold cross-validation on the training data to find the 'best' α and λ parameters, deciding to what extent features were either shrunk or removed from the model as part of the elastic net configuration.[23] We defined the 'best' α and λ parameter estimates for a model as their combined global optimum. This optimum was then defined as the most parsimonious model within 1 standard error (SE) of the model with the lowest error, as defined by the MSE. By not choosing the best performing model, we could circumvent models that might be needlessly complex and thus somewhat balance prediction accuracy and model complexity. The evaluation parameter, RMSE, for the training and internal test set was computed by taking the model MSE and computing its square root. For evaluation on external data, we had to rebuild the training model manually from the model's coefficients.[24] Occasionally, the sets of constant features differed across training and (external) test sets, requiring us to add empty columns modelling 'zero occurrence' in the test data.

Table 4 shows the baseline results for both data sets. These results are computed by using the mean of the data for prediction of every instance. The columns 'training' and 'test' refer to the 75/25 split of the data set. For the last column ('ext. test'), the two previous

[21] This was done using the *caret* package in R (Kuhn 2014).

[22] All regression models were computed using the *glmnet* package in R (Friedman *et al.* 2010), which in our opinion currently offers the most transparent and flexible implementation.

[23] The procedure followed was that outlined by Nick Sabbe:
`http://stats.stackexchange.com/questions/17609/cross-validation-with-two-parameters-elastic-net-case`
– last verified: March 2018

[24] Unfortunately, we were not able to use the *glmnet* package directly to evaluate on data other than that from the training set. It seems that training and external test data would first have to be aligned in terms of features, followed by re-computation of the model and then evaluation on external test data.

Table 4:
Baseline for both data sets

Data set	RMSE		
	training	*test*	*ext. test*
two-author set	11.1	13.0	11.5
reference set	17.4	17.0	17.3/14.1

columns are added together to be used as an external validation set: i.e. the two-author model is validated on the reference data set and vice versa. There are two baselines for the reference set: the first one was calculated over the entire set, whereas the second one was based only on those items within the same time span as the two authors. Testing the two-author model on the smaller reference sample avoids extrapolation beyond the authors' time span.

5.2 *Model results*

Based on the four feature types and four n-gram lengths, sixteen different models were computed for each data set. Table 5 shows the model results for both the reference corpus (columns 2–7) and the two-author data set (columns 8–13). The first two columns for each set show the number of constant features compared to the total number of features present for each feature type and n-gram length, giving the raw counts as well as the corresponding proportions.[25] Considering these proportions with respect to feature type and sequence length (i.e. unigram, bigram, trigram, or tetragram), one can observe several patterns with respect to the number of features extracted. For both data sets, the number of all features extracted increases with n-gram size for all four feature types. However, when considering only constant features, there is a difference for the more general character and POS types as opposed to the more specific stem and lexical types. While the general types always increase in cardinality but not in proportion in the next higher sequence, e.g. unigram to bigram, across all levels, the specific types only increase up to bigram/trigram size and then decrease again. In addition, the increase in total types is considerably higher and causes the proportion of constant types of all types to be much smaller than for the first group. This is undoubtedly due to the large number of extremely rare features, adding to the count of

[25] The number of constant features reported does not include the added variables 'author' or 'year'.

Table 5: Results for the reference data set (left) and two-author data set (right) for all four feature types, showing constant features used as input versus all features extracted and the corresponding percentage in the first two column then RMSE over training and test data as well as results for testing on the other data set ('ext.test'). 'model' lists model specifications, i.e. number of β coefficients

type-ngram	Reference data set						Two-author data set					
	input		rmse			model	input		rmse			model
	constant/total	%	*training*	*test*	*ext.test*	βs	*constant/total*	%	*training*	*test*	*ext.test*	βs
Char-1	54/69	*78*	4.5	5.1	20.9	25	37/128	*29*	11.5	12.6	17.5(14.3)	2
Char-2	914/2632	*35*	3.6	5.0	80.5	24	518/3202	*16*	11.2	12.2	17.2(14.3)	2
Char-3	7236/47156	*15*	2.9	5.2	60.2	39	2788/27631	*10*	10.0	10.7	17.9(13.8)	11
Char-4	29316/350458	*8*	3.5	2.8	35.5	315	6544/137307	*5*	10.6	12.6	17.5(13.8)	5
	constant/total	%	*training*	*test*	*ext.test*	βs	*constant/total*	%	*training*	*test*	*ext.test*	βs
POS-1	43/45	*96*	4.3	4.4	21.4	10	39/45	*87*	11.6	13.1	17.5(14.2)	0
POS-2	1219/1895	*64*	3.5	3.5	18.5	83	489/1604	*30*	10.1	8.3	17.7(15.0)	69
POS-3	10973/48673	*23*	3.3	3.5	19.2	297	1461/27802	*5*	10.2	11.2	17.8(15.0)	94
POS-4	36159/593841	*0.6*	3.3	3.8	21.0	207	1547/207858	*0.7*	10.2	12.3	17.0(14.0)	1
	constant/total	%	*training*	*test*	*ext.test*	βs	*constant/total*	%	*training*	*test*	*ext.test*	βs
Stem-1	7808/320714	*2*	3.9	3.2	21.7	45	672/38915	*2*	10.2	8.8	17.8(15.1)	53
Stem-2	36189/9589629	*0.4*	3.5	3.2	12.8	81	578/967400	*0.06*	9.9	10.2	17.6(15.6)	6
Stem-3	16613/45610366	*0.04*	4.5	3.9	15.0	208	29/3079424	*0.0009*	10.4	10.2	17.3(15.0)	7
Stem-4	2238/85402502	*0.003*	5.1	5.6	15.5	55	1/4533542	*0.00002*	11.6	13.1	17.5(14.9)	0
	constant/total	%	*training*	*test*	*ext.test*	βs	*constant/total*	%	*training*	*test*	*ext.test*	βs
Lex-1	13782/741069	*2*	2.8	3.0	19.0	25	579/101630	*0.6*	10.3	9.3	18.0(14.3)	25
Lex-2	35773/11790813	*0.3*	2.9	2.2	20.6	78	633/1147540	*0.06*	10.6	10.2	17.9(14.5)	100
Lex-3	21515/47085085	*0.05*	3.4	2.4	17.0	183	78/3226493	*0.002*	11.0	9.4	17.6(16.9)	10
Lex-4	4811/89673339	*0.005*	4.3	3.4	na	35	0/4911004	na	na	na	na	na

total but not constant features. These patterns are primarily observable in the two-author data set, and are a little less pronounced for the reference data set. The remaining four columns for each set show training, test, and external test set RMSE, and the complexity of the model measured by the count of β coefficients.[26]

5.2.1 Reference corpus

We first consider models specific to the reference corpus, noting baseline results of 17.4 (training), 17.0 (test) and 11.5 (external test), as shown in Table 4. From the results in Table 5, one can observe that for character n-grams, model accuracy ranges from 2.9 to 4.5 years for the training set and from 2.8 to 5.2 years for the test set. Models 'Char-2' and 'Char-3' are best at balancing accuracy of prediction and model parsimony. With an RMSE of 20.9, 'Char-1' performs best on the two-author data, although this is still far from the baseline of 11.5, with the other three models being even less accurate (RMSE: 35–80). This suggests that there is little similarity between the data sets with regard to character n-grams. The results for the syntactic sequences (POS-n) are very regular over all four n-gram sizes, varying between an RMSE of 3.3–4.3 years for the training set and 3.5–4.4 years for the test set. External validation error on the two-author data set is lower than for the character n-grams but still not comparable with the baseline (18.5–21.4). Model complexity increases noticeably with n-gram size: our 'POS-1' model achieves an accuracy of 4.3 on the training set and 4.4 on the test set. While the bigram model 'POS-2' decreases this to 3.5 for both sets, it also adds 73 more predictors. Similarly, 'POS-3' and 'POS-4' both obtain an RMSE of 3.3 on the training set, but use 297 and 207 predictors, respectively. The word stem unigram and bigram models perform slightly better than their POS counterparts, with model accuracy slightly deteriorating after that, despite using more predictors. 'Stem-1' and 'Stem-2' achieve 3.9 and 3.5 on the training set, with 3.2 for both on the test set. This deteriorates to 4.5 and 3.9 for 'Stem-3' and then to 5.1 and 5.6 for 'Stem-4'. External validation is better than for the two previous types (12.8–21.7), but still cannot quite compete with the baseline. Overall, syntactic

[26] The coefficient count β does not include the intercept.

word features (Lex-n) and 'Lex-1', and 'Lex-2' in particular, yield the most accurate models. The unigram and bigram models obtain an error of 2.8–2.9 on the training set and 2.2–3.0 on the test set. 'Lex-1' might be considered the best model overall, as it has 53 fewer predictors than 'Lex-2', yet performs only slightly less well on the training and test sets (0.1 and 0.8 years, respectively). The external validation error (17–20.6) is higher than for stem n-grams, indicating that the two data sets might be 'closest' for that type. As previously noted, some of the above models seem rather complex and, given the tendency of elastic nets to select correlated predictors, poses the question of whether so much complexity is needed to achieve model accuracy.

In order to see which models have a large number of correlated predictors, we consider the corresponding uncorrelated models by rerunning the same experiments, but using only the lasso method, i.e. setting α to 1. This highlights several aspects of the regression models computed earlier: a simple model of ~10–30 predictors can still be improved by adding features, in the sense that these contribute enough new information to improve prediction accuracy. In most cases, however, adding more features to a model of 80 predictors rarely improves prediction accuracy. Compare adding 7 features to achieve a −0.3/−0.5 error decrease ('Lex-4') to adding 151 features for a −0.5/−0.5 RMSE decrease ('Lex-3') for training set and test set respectively. What is also notable is that most lower n-gram models do not have any correlated predictors, seeing that elastic net and lasso methods yield the same models, whereas the number of correlated predictors rises with n-gram size up to trigram size, whereafter model size suddenly decreases more or less dramatically.[27] This strongly suggests that there is most overlap for trigram models on the most changing features used in each time slice. Thus, while there is likely to be most background language change in syntactic word features, all types produce accurate enough models to suggest that reasonably interesting temporal change must have taken place. The language change aspect is examined in more detail in Section 6.

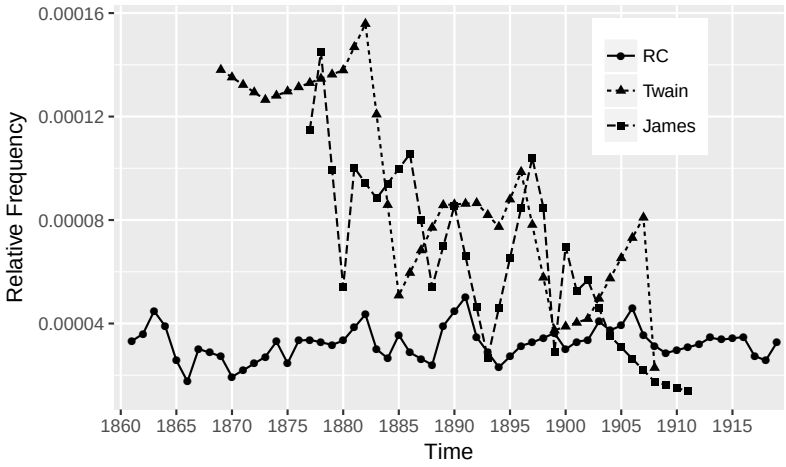
[27] This is with the exception of character n-grams, as these would probably need to grow to average word length in order to be less correlated.

5.2.2 Two-author data set

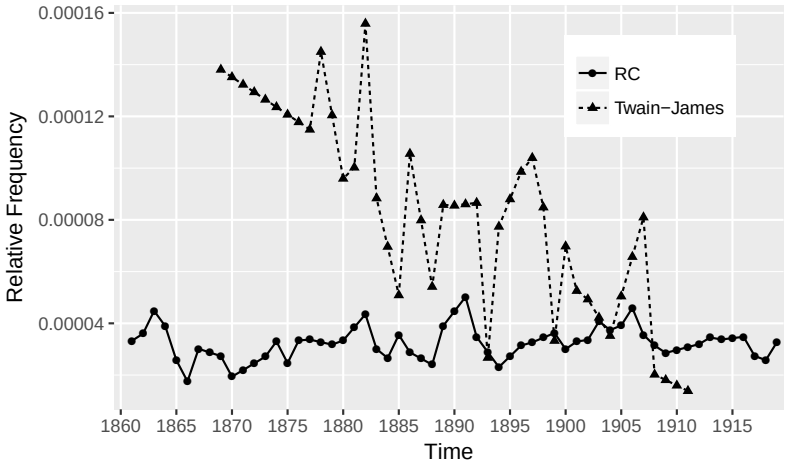
We now turn to the models intended to capture individual change, specifically in James' or Twain's language. The baseline results for the two authors yielded 11.1 (training), 13.0 (test) and 17.3/14.1 (external test). Beginning with the character n-gram models, Table 5 shows that 'Char-1' and 'Char-2' are very close to the baseline, containing very few predictors, indicating that these two types carried little discriminatory power. The trigram model 'Char-3' is the best character model, with 10/10.7 RMSE for training and test set, where the error is much lower than the baseline of 13, especially for the test set. The 'Char-4' model does not quite reach the same accuracy, although it is an improvement on the first two models. The results on the external test data are consistently congruent with the baseline for that set. Moving on to syntactic sequences, the unigram model 'POS-1' is actually the null model, as it is the most parsimonious model within one standard error of the best model with 38 features, suggesting that this type is not discriminatory enough in relation to publication year. The best POS model is 'POS-2' with 10.2/8.3 on training and test set respectively, but it increases complexity by adding 69 predictors. 'POS-3' adds even more complexity (94 predictors), but performs worse than 'POS-2'. Interestingly, the 94 predictors in 'POS-3' have the same predictive power on the training set as 'POS-4's one and only predictor ⟨VBD VBN IN JJ⟩.[28]

Figure 1 depicts the tetragram ⟨VBD VBN IN JJ⟩ for Twain and James individually (Figure 1a) and combined together (Figure 1b). Even though relative frequency values vary over only a small range (0.00004–0.00016) for both James and Twain (Figure 1a), there is a discernible downward trend over time, offering a fair indication of temporal origin. The combined plot, though a generalization, still presents a fair approximation of each individual plot. In comparison, the same feature exhibits less of a trend over time for the reference corpus. The prediction accuracy of stem models is comparable to that of character and POS n-grams, while models tend to be more parsimonious. Results range from 9.9–10.4 on the training set and 8.8–

[28] This tag represents a sequence of ⟨a verb in past tense (VBD), a verb in past participle (VBN), a preposition (IN) and an adjective (JJ)⟩ as in ⟨were.VBD accompanied.VBN by.IN restless.JJ⟩.



(a) ⟨VBD VBN IN JJ⟩ for James, Twain, and the RC

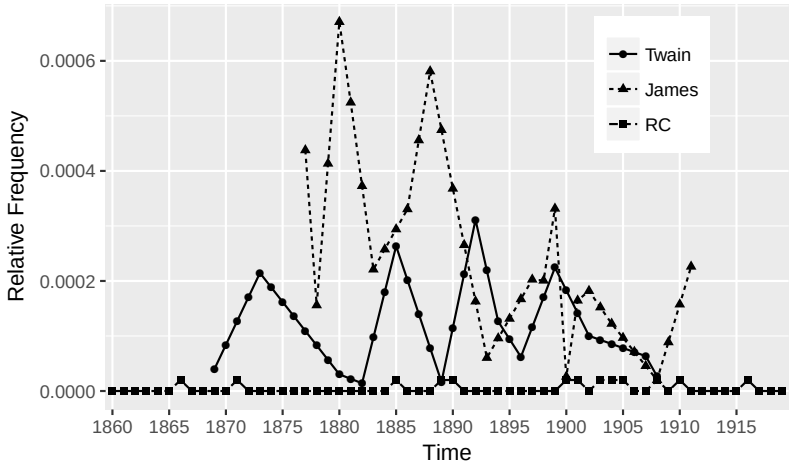


(b) ⟨VBD VBN IN JJ⟩ for James + Twain, and the RC

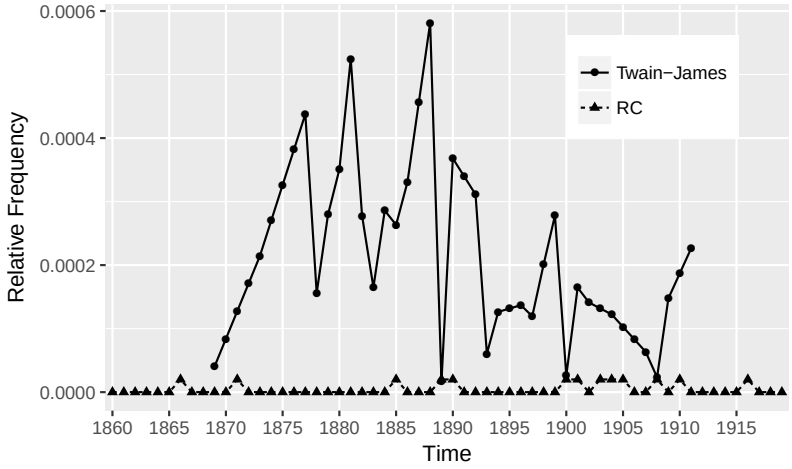
Figure 1: The development over time of the tetragram ⟨VBD VBN IN JJ⟩, showing relative frequency in relation to all tokens for the reference corpus (RC) and for the two authors. Figure (a) shows the feature for Twain and James separately. Figure (b) shows the combined two-author frequency, averaged for years when both published work

10.2 on the test set for 'Stem-1', 'Stem-2' and 'Stem-3'. For word stem tetragrams, the number of constant features drops to one (which is the feature ⟨i don t know⟩), causing the null model to be selected.[29] Figure 2 depicts this feature for Twain and James separately (Figure 2a) and combined into one (Figure 2b), each time alongside the reference corpus. Variability somewhat decreases over time for the two

[29] The corresponding syntactic word feature would be: ⟨i do n't know⟩.



(a) ⟨i don t know⟩ for James, Twain, and the RC



(b) ⟨i don t know⟩ for James + Twain, and the RC

Figure 2: The stem feature ⟨i don t know⟩ for the reference corpus, and for Twain and James separately (Figure (a)) and combined (Figure (b))

authors, if less markedly than in the previous case, and while there is a downward trend for James, there is no specific trend visible for Twain. Combining their two plots over time yields a less appropriate approximation to each individual, indicating that there are stronger differences between them. Interestingly, this tetragram feature was not constant over the reference corpus, in spite of a much larger data selection available – its line in the plots indicates occurrence rather than relative frequency in both Figures 2a and 2b. When the feature

occurs, the raw count generally varies between 1 and 2 and never exceeds 6 (total token count for the same year is 2,228,655). This indicates a very different usage from James and Twain, and could imply that other synonymous forms were more common, e.g. 'I do not know' or that first person references were used less frequently than by the two authors. Examining alternative, high-ranking models for 'Stem-4' yields a pairing of ⟨i don t know⟩ with the 'author' feature. Figure 2 shows that relative frequencies for James and Twain are reasonably different until 1890, with little overlap, possibly rendering separation by authorial source more useful than in the previous cases.

This result shows that, although this feature was used by both James and Twain, it was rare in general language at the time. James initially used it more than Twain, but, over time, their rates of use appear closer. Thus, there are two different dimensions to this analysis, the constancy of a feature over a corpus, and its relative frequency. The main difference between the reference corpus and the two-author data set is that of constancy, whereas the main difference between Twain and James pertains to the feature's relative frequency. In any case, a more detailed investigation is needed to exclude possible confounding factors, such as genre or narrative perspective, to confirm that this pattern is rooted in stylistic differences only.

Finally, we consider the most specific linguistic type, syntactic word features. The best overall models are 'Lex-1' and 'Lex-3', with 10.3/11 on the training set and 9.3/9.4 on the test set. 'Lex-2' is more complex (100 predictors) and yet a little less accurate.

These results suggest that the more general feature types (character/POS) need longer sequences to be discriminative. In contrast, stem n-grams are fairly accurate, sometimes even with only very few predictors, provided there are enough input features. The fact that the 'author' variable was never chosen to be a part of any model suggests either that Twain and James are rather similar with respect to their shared constant features that are discriminatory over time, or that their rate of change is entirely different, making a distinction for the level not helpful.

5.3 *Comparison with previous results*

The final part of the experiments is to compare these results with those from our previous study on syntactic word unigrams (Klauss-

Table 6:
Results for previous work (Klaussner and Vogel 2015), showing RMSE and model size for the reference corpus, the James and Twain data set, and the combination of all three data sets

Reference set				
Model	*training*	*test*	*ext.test*	βs
1	3.2	4	15.4(T)/20.3(J)	4
2	11.9	12.1	42.2(T)/44.7(J)	5
Two-author set				
Model	*training*	*test*	*ext.test*	βs
1	5.5	7.2	–	5
2	5.2	8	–	7
Combined set				
Model	*training*	*test*	*ext.test*	βs
1	2.8	1.8	–	5

ner and Vogel 2015). Table 6 shows the results for the reference corpus, the two-author data set, and a third corpus combining all data sets in one. In comparison to earlier experiments, our results for the reference corpus add ~1 year accuracy in prediction. The results for the two-author data set are less accurate. This confirms that taking only constant features for prediction and discarding all others results in the loss of valuable predictors. In part this could be due to a feature's non-occurrence in particular years, possibly aiding the statistical technique to discriminate more easily between years. Using features occurring less reliably has to be applied with caution as, on the very infrequent side of the frequency spectrum, there lurks statistical optimization, which would not only yield unstable models, but would also focus less on characteristic and more on idiosyncratic aspects of the particular data set under study. One therefore needs to differentiate between features that are infrequent during an author's lifetime, but very frequent in those years when they do occur, and features that are consistently infrequent. An extreme case of this would be sets of *hapax legomena*. The reason why the models are more accurate for frequent, but not quite constant, features may be that authors are likely to be more consistent for features that they use constantly throughout their literary career, than for those that they use less regularly. In any case, we emphasize that our purpose is not achieving the highest possible accuracy in assignment of temporal provenance, but in understanding what fea-

tures change in frequency over time, and how those changes are to be interpreted. The latter task is open-ended, but depends on the former.

6 ANALYSIS OF LANGUAGE CHANGE

In this section, we consider salient features of the regression models presented in Section 5.2. In order to select those features that change most over time, we rank the respective model's predictors according to the absolute weight it is assigned in the model, thereby selecting features that increase and decrease linearly over time. However, to identify features that did not exhibit any change over time, we had to exclude features that rated high on either linear or non-linear change. For this purpose, we evaluated all features separately with respect to the response variable, and selected those that rated low on both linear and non-linear relationships. Section 6.1 introduces some general language change trends and Section 6.2 then analyses the data for the two authors in comparison with the reference corpus.

6.1 *Reference language change*

In the following, we present some aspects of general language change based on the changes detected in the reference corpus. This is not presented as an exhaustive list, but merely as a series of examples. In the following, we focus on lexical and syntactic change.

Figure 3 shows samples of the highest-rated features for each of the three categories: 'increase over time', 'decrease over time' and 'no change'. Considering shorter n-gram sizes shows that there might be considerable overlap between different models of the same feature type but different n-gram size, and also between different feature types. Figure 4 shows the word n-gram ⟨a matter of fact⟩ and its hypergram ⟨a matter of⟩. As can be seen from the difference in frequency, there are a number of other frequent realizations of ⟨a matter of⟩, such as ⟨a matter of concern⟩ or ⟨a matter of urgency⟩. There are cases where the more specific sequence accounts for most of the occurrences of the generic one, whereas in cases like these it only accounts for part of them.

Figure 5 shows the most prominent syntactic tetragrams. The sequences ⟨DT NN IN WRB⟩ and ⟨DT NN TO VBG⟩ both increase over

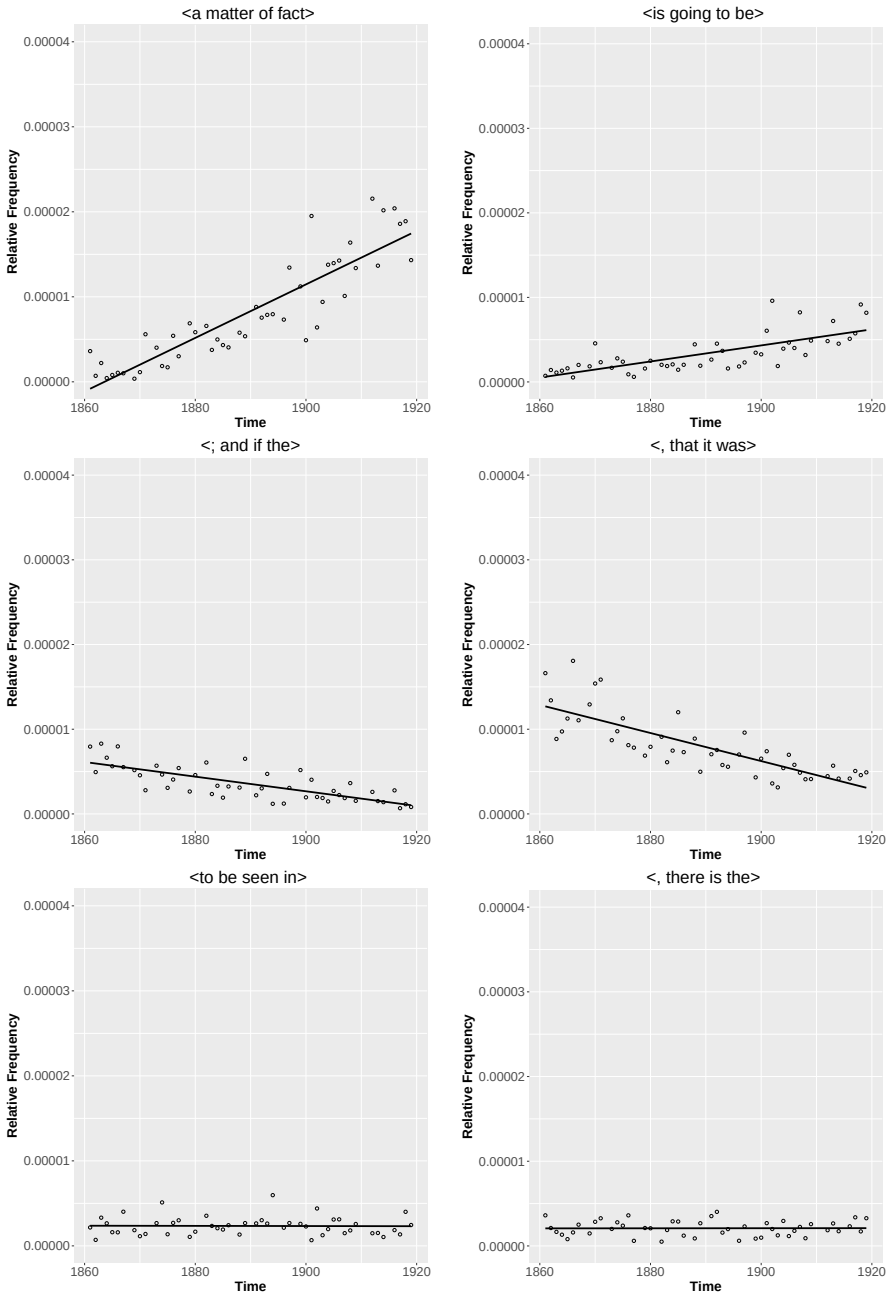


Figure 3: Reference corpus: relative frequency of several syntactic word tetragrams, exhibiting ‘increase’, ‘decrease’, or ‘no perceptible change’ over time

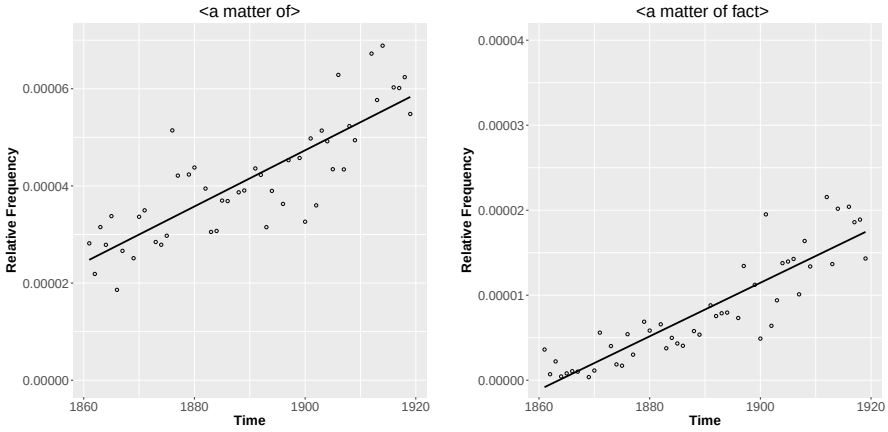


Figure 4: Reference corpus: relative frequency for ⟨a matter of⟩ and ⟨a matter of fact⟩

time. Phrases such as ⟨the fact that when⟩ or ⟨the secret of where⟩ are examples of the former, and ⟨no objection to saying/taking⟩ or ⟨a view to showing/discovering⟩ are examples of the latter. Thus, depending on whether the words in the sequence are content or function words, and whether they are part of a collocation, certain combinations will be more frequent (⟨a view to⟩/⟨no objection to⟩), while others may be more variable. The shorter variant of this ⟨DT NN TO⟩ does not seem to be discriminative over time. Similarly, examining some corresponding syntactic word sequences ⟨a.DT view.NN to.TO⟩ and ⟨no.DT objection.NN to.TO⟩ shows that, although constant, they do appear to change in a rather random fashion. The more specific tetragram sequences, such as ⟨no objection to saying⟩ are usually not constant. Realizations of decreasing POS features (⟨CC NN VBP PP⟩ and ⟨IN VBG , IN⟩), also yield patterns of fixed and varying units: ⟨and pride/happiness attend her⟩ and ⟨by saying, that⟩/⟨without murmuring, because⟩. The syntactic combinations that show the least development during this time span are ⟨EX VBZ RB JJR⟩ with examples such as ⟨there is far more/less⟩/⟨there's something stronger⟩, and ⟨VBD NN DT NN⟩ with examples like ⟨was nothing the matter⟩ or ⟨made music all day⟩.

Given the size of the corpus, one would expect a variety of feature realizations to be among the constant features, especially in the presence of multiple genres, and the differences in language us-

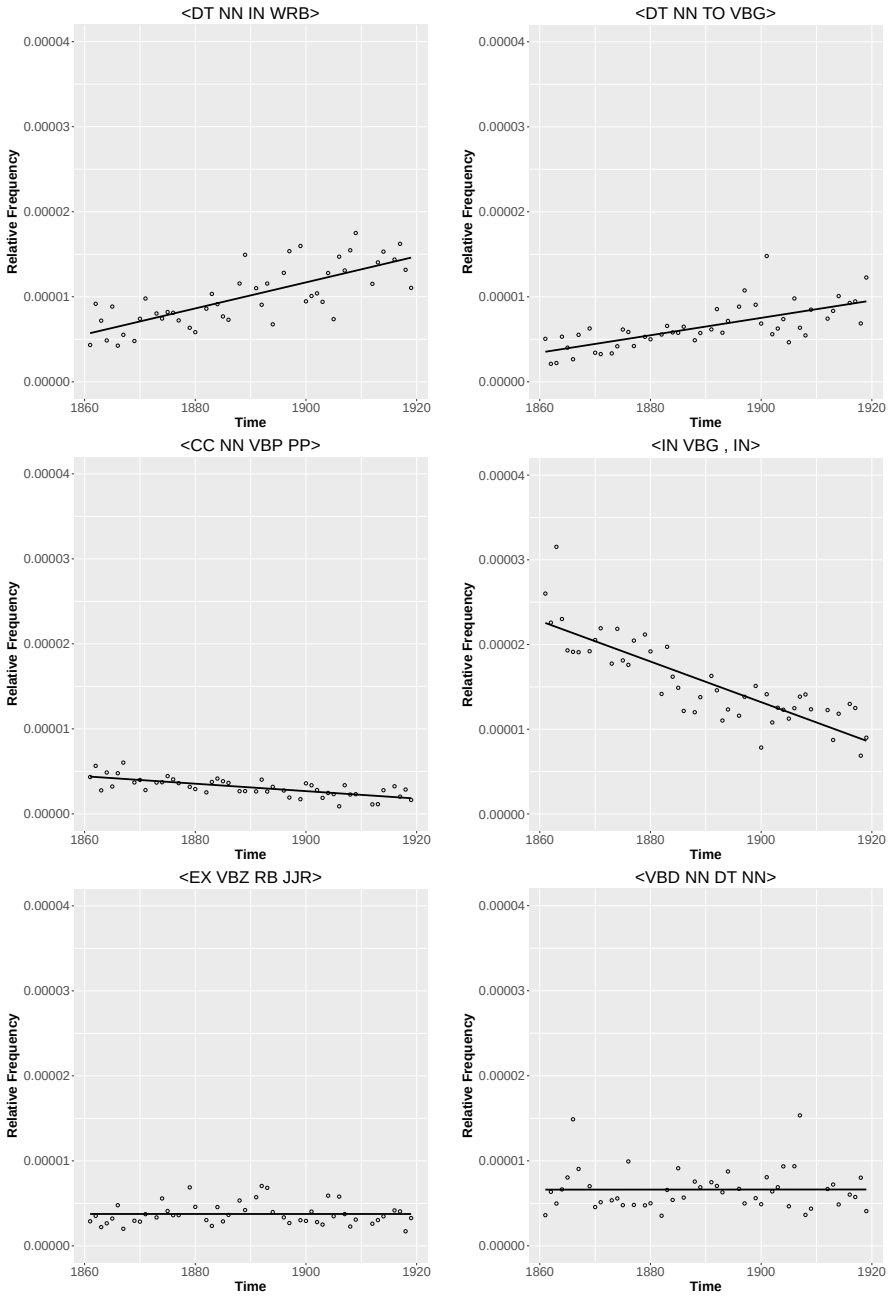


Figure 5: Reference corpus: relative frequency of several syntactic tetragrams, exhibiting ‘increase’, ‘decrease’, or ‘no perceptible change’ over time

age found in these genres. In spite of this, most of the consistent features or their generalizations present here seem to be expressing opinions, or to be ways of organizing these, such as ⟨a matter of fact⟩ or ⟨a view to⟩/⟨no objection to⟩, which are items that could be expected to appear in a variety of contexts. In order to identify change that is not general to all written language, one might investigate change in different genres, such as fiction, or newspaper articles. The most dramatic change is found in very general POS n-grams, which incidentally also display more spread. In contrast to syntactic word n-grams, POS n-grams are more volatile in that they represent a group of words that could possibly change or give rise to different frequencies.

6.2 *Two-author language change*

We now turn to the analysis of the two authors, to examine how their language changed or stayed the same over time, while also taking into consideration how their language differed from the reference language of the time. In the following, we consider different aspects of how style could vary. Section 6.2.1 considers differences between constant feature sets of lexical types. Sections 6.2.2 and 6.2.3 consider stylistic differences between the reference corpus and the two authors, and then any stylistic differences between the two authors.

6.2.1 Constant features

In order to explore the stylistic differences between Mark Twain and Henry James, we examine different sets of constant terms: those they share and those they do not share. It is important to note that constancy does not necessarily imply high frequency, and that one word or expression could be constant for only one author but more frequent overall for the other.

Figure 6 shows 'wordclouds' based on their individual non-shared noun, interrogative pronoun, and adjective type features. We grouped these together for inspection since they could all occur in noun phrases but, unlike pronouns and determiners, are less grammatically controlled, and therefore more meaningful.

Table 7 shows the relative frequency data for wordcloud items, ordered by relative frequency, showing the median rank of each item in the wordcloud group, and among all constant features for that author.

Table 7: Relative frequencies and rank for the 20 most frequent wordcloud items, in the works of Twain, of James, and of both authors together. Words are listed by relative frequency rank (RFR). The ‘wcr’ columns show wordcloud group ranking. The ‘cr’ columns show rank among all constant items

RFR	Individual items								Shared items							
	Twain	*RF*	*wcr*	*cr*	*James*	*RF*	*wcr*	cr	*Twain*	*RF*	*wcr*	*cr*	*James*	*RF*	*wcr*	*cr*
1	god	0.00039	10	344	mr	0.00162	1	112	what	0.00239	1	63	what	0.00383	1	38
2	money	0.00037	1	266	lady	0.00053	2	206	time	0.00184	2	72	little	0.00177	2	65
3	boy	0.00037	4	316	de	0.00047	168	722	man	0.00145	3	79	who	0.00146	4	76
4	mother	0.00033	7	348	father	0.00046	23	399	other	0.0014	5	78	time	0.00134	3	80
5	everybody	0.00031	2	264	whom	0.00045	3	207	good	0.00132	4	87	great	0.00119	7	89
6	body	0.00029	3	276	lord	0.00040	134	733	who	0.0013	6	80	other	0.00117	5	102
7	sir	0.00028	19	381	dear	0.00034	4	260	way	0.00129	7	88	young	0.00107	14	117
8	boys	0.00028	17	385	companion	0.00034	5	242	old	0.00113	11	104	way	0.00104	9	103
9	dead	0.00027	6	275	charming	0.00029	17	352	little	0.00112	10	103	moment	0.00104	6	99
10	door	0.00027	5	319	effect	0.00028	7	273	thing	0.00105	9	106	man	0.00101	8	101
11	family	0.00027	9	311	particular	0.00027	10	305	day	0.00097	8	100	good	0.001	10	110
12	children	0.00024	14	330	view	0.00027	12	334	people	0.00096	12	116	nothing	0.00099	11	105
13	ready	0.00023	8	333	possible	0.00026	8	287	great	0.00082	13	121	more	0.00095	12	114
14	anybody	0.00023	12	409	reason	0.00026	9	318	nothing	0.00079	14	134	things	0.00094	15	118
15	nobody	0.00021	13	375	round	0.00025	6	291	more	0.00073	15	136	own	0.00091	16	123
16	week	0.00021	16	400	impression	0.00024	15	327	place	0.00067	18	158	old	0.00088	19	127
17	river	0.00021	22	426	tone	0.00023	13	321	last	0.00067	16	163	something	0.00082	13	119
18	girl	0.00021	20	441	rate	0.00023	11	319	things	0.00065	19	159	thing	0.00081	20	131
19	minutes	0.0002	23	447	love	0.00022	25	396	years	0.00064	17	150	last	0.00076	18	130
20	bed	0.0002	18	436	bad	0.00022	14	322	night	0.00063	20	155	eyes	0.00074	17	128

Figure 6: Noun, interrogative pronoun, and adjective type wordclouds for Twain (left) and James (right), based on non-shared constant features

Twain's most prominent words express existential concepts, apparently pertaining to a more questioning nature, e.g. 'god', 'money', ' everybody', 'anybody', 'nobody', 'family', 'mother', 'children', 'dead', 'heaven', 'church', 'trial', and 'soul'. In contrast, James' most prominent words in this group are more prosaic, e.g. 'mr', 'father', 'lady', 'dear', 'whom', 'lord', 'charming', 'companion', 'impression'.[30] It is interesting to note the difference between James' most frequently used form of address, 'Mr', and Twain's 'Sir' – 'Mr' suggests that one could address both a superior and an equal, whereas 'Sir' is used predominantly when addressing a superior, which is plausible as Twain also wrote about less wealthy people.[31] James' list also includes the French word 'de', often found in names and addresses and, which was incorrectly tagged here as a proper noun.[32] There are some other interesting contrasts, such as 'conscience', which is constant for Twain, and 'conscious'/'consciousness', constant for James. Twain's words suggest more intense situations, intimating both good and bad, e.g. 'crime', 'cruel', 'blood', 'dark', 'lonely', 'alive', 'peace'. James' most negative words in this group are 'sad', 'helpless', 'victim', indicating that Twain's language was more explicit. While James' stories do contain conflicts, they were possibly more veiled than in Twain's texts.

[30] As all data was transformed to lowercase for analysis, words, such as 'Mr' appear that way in figures as well.

[31] The word 'Sir' is ranked 19 among wordcloud features and 381 among Twain's constant features.

[32] The word 'de' is ranked 168 among wordcloud features and 722 among James' constant features.

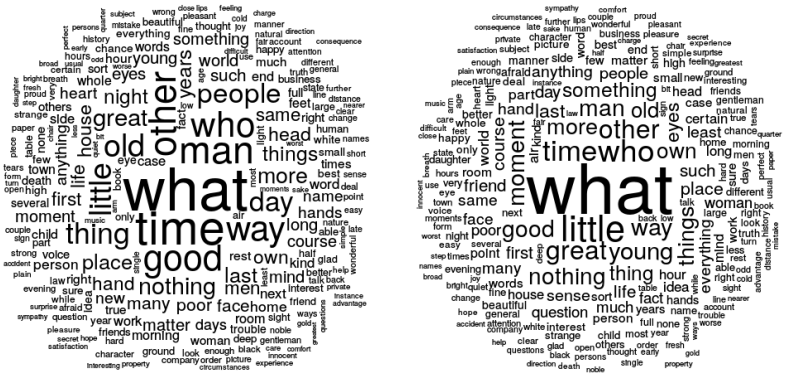


Figure 7: Noun and adjective wordclouds for Twain (left) and James (right), based on their shared constant features

Figure 7 shows the wordclouds for their shared constant nouns, interrogative pronouns, and adjectives. Their most prominent words are quite similar here, e.g. ‘what’, ‘time’, ‘little’, ‘good’, and ‘young’. There are some less frequent words for both that are interesting to consider, with a wider semantic range: ‘circumstances’, ‘feeling’, ‘consequence’, ‘believe’, ‘truth’, and ‘pleasure’. Depending on context, these words might take on either a more superficial or deeper meaning, e.g. ‘I believe you’re right’ and ‘I believe in one Christ’.

Interestingly, both authors took an avid interest in history, evidenced by the syntactic unigram ⟨history⟩ being among their shared constant features. Both Blair (1963) and Thomas M. Walsh and Thomas D. Zlatic (1981) note that history played an important part in Twain’s personal as well as his professional life, even if he did not always incorporate his knowledge consistently into his works (Williams 1965). In his 1884 essay ‘The Art of Fiction’, James actually claims his place among historians, since a novelist chronicles life, and as ‘picture is reality, so the novel is history’ (James 1884). All of the two authors’ constant word unigrams are present in the constant features of the reference corpus, except for James’ term ‘vagueness’.[33]

While constant word unigrams reveal a great deal about recurring concepts, longer sequences might hold more information about unique aspects of style, as these tend to be more generic. Table 8 shows examples of constant bigram and trigram word sequences and

[33] Although using wordclouds can give some insight into the data, it cannot replace the study of actual word frequency distributions. The extended set of constant features can be found here: `www.scss.tcd.ie/clg/4thIWCH/`.

Table 8: Examples of constant syntactic word sequences (bigrams and trigrams) characteristic of: both authors (columns 3–7); James alone (columns 8–10); Twain alone (columns 11–13). The column 'cr' indicates median rank, considering bigram and trigrams separately. For readability, relative frequencies are multiplied by 100. In the interest of space, ⟨...⟩ are omitted here

Group	Type	shared bigrams/trigrams					James only			Twain only		
		n-gram	*RF(J)*	*cr(J)*	*RF(T)*	*cr(T)*	*n-gram*	*RF*	*cr*	*n-gram*	*RF*	*cr*
Possessives	n2-IN-m	by his	0.003	589	0.005	589	as his	0.005	795	into his	0.012	1069
	n2-IN-f	with her	0.007	130	0.005	547	on her	0.029	301	–	–	–
Body parts	n2-sg	hand ,	0.004	428	0.011	515	eye ,	0.025	1116	head ,	0.093	665
	n2-pl	eyes ,	0.009	446	0.046	815	hands ,	0.051	1016	–	–	–
	n2-PP$-m	his eyes	0.003	414	0.003	822	–	–	–	his eye	0.035	995
	n2-PP$-f	–	–	–	–	–	her head	0.039	431	her face	0.01	824
Expressions	n3	a matter of	0.004	158	0.046	284	in spite of	0.016	24	on account of	0.007	169
Consciousness	n2	i know	0.008	198	0.007	225	i mean	0.06	258	i knew	0.008	569
	n3	, i think	0.012	185	0.008	307	i think ,	0.005	154	. i know	0.007	233
Existential	n2	there is	0.035	296	0.048	157	there are	0.01	467	there 's	0.004	487
	n3	. there was	0.004	40	0.004	33	–	–		, and there	0.004	84

their frequencies found in the data for Twain, for James, and for Twain and James together. These lists are mutually exclusive, meaning that each term is shown only once, in the set where it is most constantly used. The rows group together n-grams by selection category. The first group contains bigram sequences of a noun followed by a preposition followed by either a male or female possessive pronoun. The second group contains singular or plural body references, either followed by a comma, or preceded by a male or female possessive pronoun. The third group contains expressions that are used for emphasis or contrast. The last two groups focus on items expressing some epistemic commitment, or with an existential construction.

Twain's language, in particular, abounds with a great variety of body references, some of which are also used by James. However, James tends to focus on body descriptions, e.g. 'face', 'eyes', 'hands', whereas Twain's constant terms include items used more abstractly, such as 'heart'. Twain's language also features many more 'existential' constructions, such as ⟨there 's⟩, which are also found in James, but with less variety. Both authors use expressions indicating reflection or thought (⟨I know⟩, ⟨I think⟩, etc.). Twain's constant terms also include the expression ⟨don't know⟩, which James does not appear to use. James seems to use contrasting features more often, e.g. ⟨in spite of⟩ or ⟨, however ,⟩, which Twain appears to employ more sparingly. Both use the male perspective more than the female one, i.e. their constant feature lists both contain various possessive and regular pronoun constructions for male characters, which are not present in the same quantity for female characters.

However, in order to properly verify these impressions, one needs to take a closer look at the actual number of constructions in each group, and their respective frequencies. We begin by considering constructions containing existential 'there' and its overall unigram relative frequency in all three corpora; the corresponding plot is shown in Figure 8. On average, Twain's usage is a little higher (ca. 0.002) than that of James and of the reference corpus, which are both around 0.0018. Table 9 shows details about the number of types for a particular item, for instance in what constructions the feature ⟨there.EX⟩ appears. This shows that Twain clearly has more constant existential types than James and, as the frequency analysis showed, he also uses

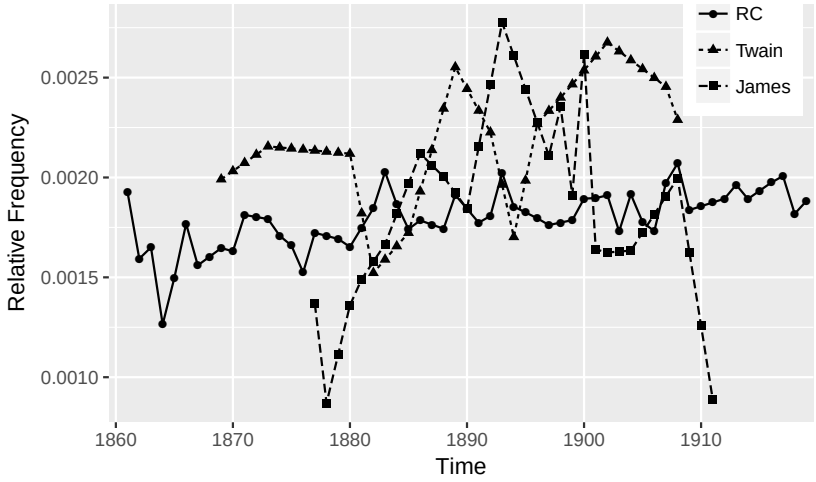


Figure 8: Existential ⟨there⟩ for all three corpora

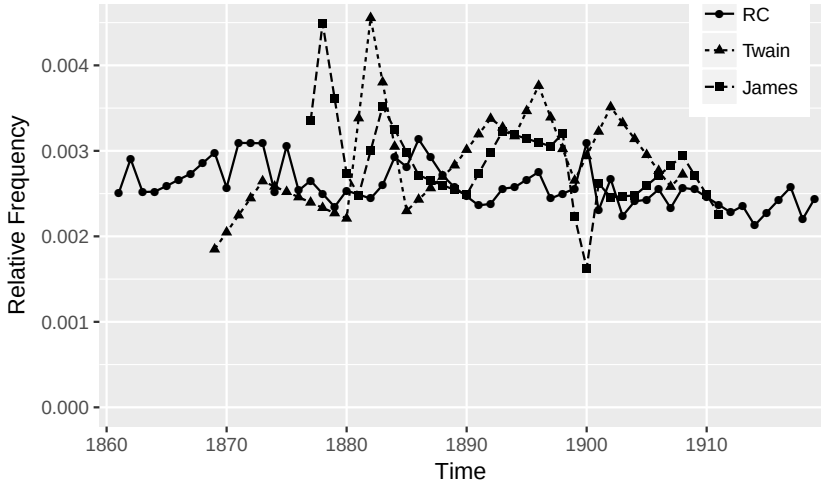
Table 9: Frequency and number of feature types for prominent constructions

	James			Twain			Shared	
Variable	*1-gram:μ*	*Lex2*	*Lex3*	*1-gram:μ*	*Lex2*	*Lex3*	*Lex2*	*Lex3*
⟨*there.EX*⟩	0.0018	3	–	0.002	7	4	4	2
body parts sing	0.0028	12	4	0.0029	14	2	4	–
body parts pl	0.001	1	–	0.001	5	1	3	–
female pr	0.023	8	1	0.008	20	5	17	–
male pr	0.025	9	2	0.025	49	54	27	9

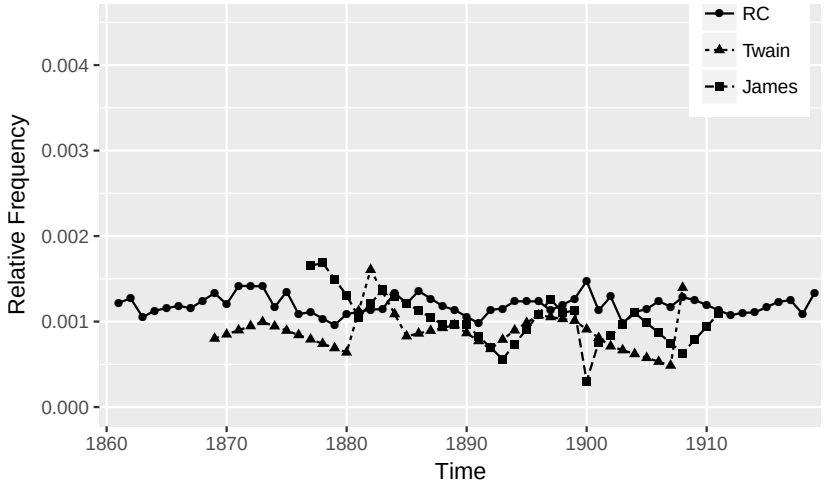
them more often. There is also an increase in usage over time for both authors, as well as for the reference corpus.

Figure 9 depicts frequency rates for body references: Figures 9a and 9b show singular and plural body parts, respectively. Interestingly, average use for body references lies above the reference corpus for singular items and below it for plural items, in both Twain and James.[34] There seems to be a decrease in usage for both types over time, with a more dramatic decrease for plural body parts. The difference between the two authors lies primarily in the variety of constructions used: there tends to be more variety in Twain's constant features – this does not mean that James does not use these features at all, but that there are fewer features that James uses regularly.

[34] The frequency rates for the reference set are 0.0026 and 0.0012, respectively.



(a) Frequency of singular body parts for James, Twain, and the RC

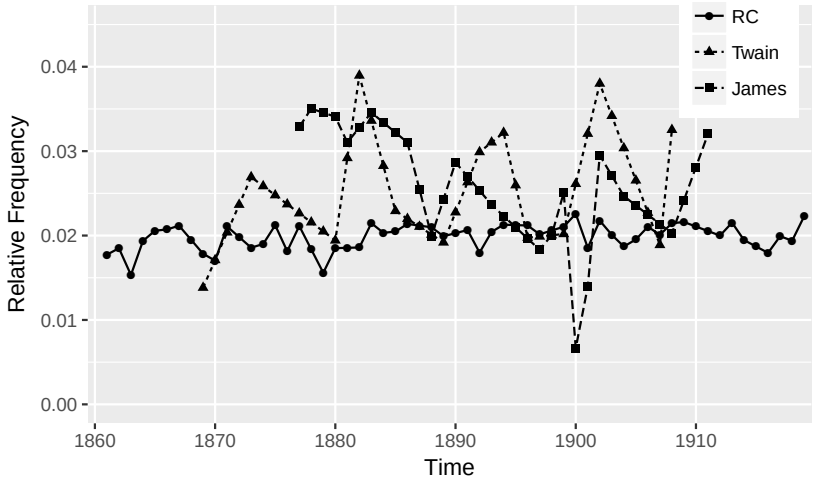


(b) Frequency of plural body parts for James, Twain, and the RC

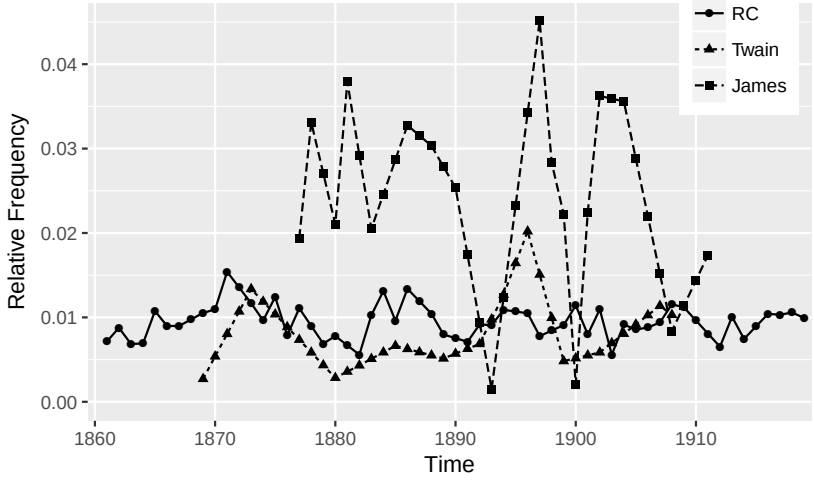
Figure 9: Body part constructions for all three corpora

Figure 10 shows the frequency rates for possessive and regular pronouns, with masculine forms in Figure 10a and feminine forms in Figure 10b. Both authors use the male perspective much more than was usual for the time, compared with the average rate of 0.025 to 0.02 in the reference corpus. Furthermore, James (0.023) refers to women through female pronouns more than twice as much as Twain (0.008), or the reference corpus (0.009). Incidentally James' constant bigram list also includes ⟨woman ,⟩ and ⟨women ,⟩ – it thus appears

Figure 10: Male and female references for all three corpora



(a) Frequency of male pronouns for James, Twain, and the RC



(b) Frequency of female pronouns for James, Twain, and the RC

as though James focused his narrative on women much more than was usual for his time. In contrast, Twain has markedly more varied constant constructions featuring pronoun references, especially for males. This could mean one of two things: either that he is more variable in his language describing people, given that he has more common phrases, or in fact that he is less variable, as he tends to draw more often from a limited set. Without a comparison with more contem-

poraneous authors, to examine the proportion of gendered pronoun constructions in their non-constant bigrams, it is not clear whether this aspect is usual or unusual. For instance, James might only have a few constant constructions, changing his language use depending on the situation.

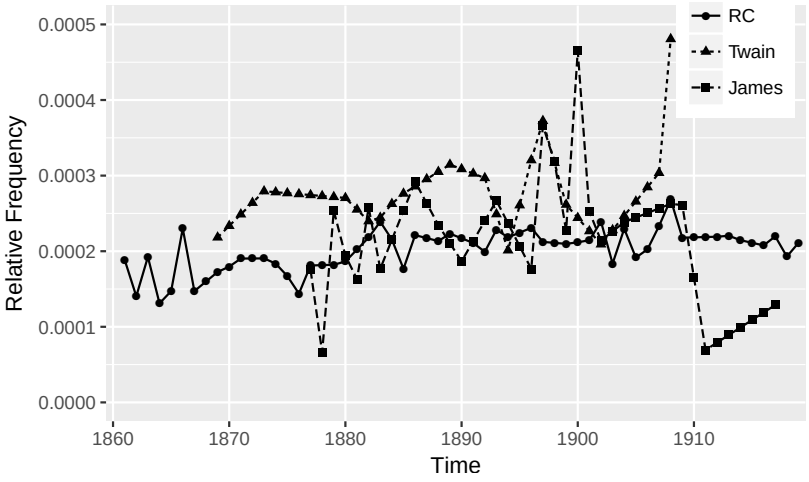
6.2.2 Stylistic differences with the reference set

In order to explore any differences from the reference language, we consider the shared salient features, i.e. the features that appear in Twain, in James, and in the reference corpus. Among the character n-gram models, there are no common predictors, except for the letter ⟨q⟩ in the unigram model. All models have a positive weight for this predictor, but only the authors show a clear upward trend over time. All word stem and syntactic word n-gram models yield one shared bigram ⟨, by⟩, which is shown in Figure 11 and Figure 12, together with three highly weighted shared POS bigrams ⟨CC EX⟩, ⟨WDT ,⟩ and ⟨MD ,⟩.

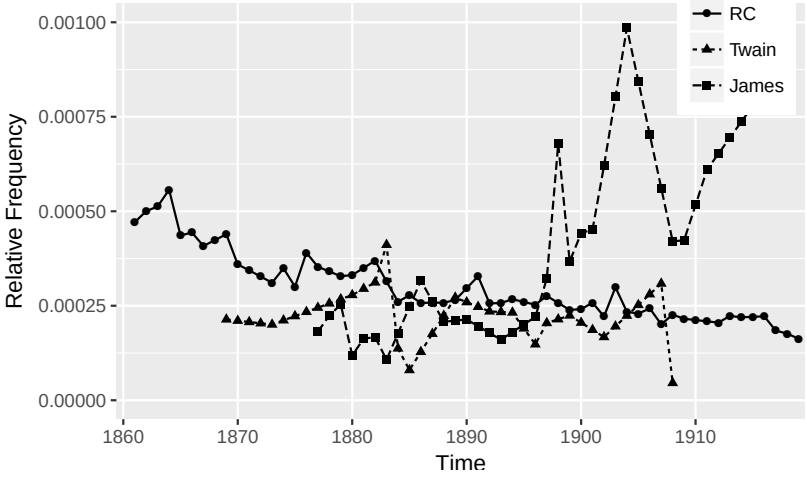
The bigram ⟨CC EX⟩ realizes expressions such as ⟨but there⟩ or ⟨and there⟩, that have already been mentioned earlier with respect to the constant features in James and Twain. For this POS bigram, their average rate tends to be higher than that of the reference corpus. What is noticeable for the other three features is that the three data sets are rather well separated, with James having the highest usage of all. This will be explored in more depth as part of the between-author analysis in Section 6.2.3. All lines show some development over time, explaining why these are salient features in the models.

With respect to syntactic changes, there seems to be a marked reduction in noun phrase constructions for the two authors, a reduction that is not present in the reference corpus, as shown for two examples in Figure 13. This trend can also be observed in several other noun-phrase-based sequences, such as ⟨DT NN NN⟩, ⟨IN NN NNS⟩, ⟨JJ NN NNS⟩, ⟨NN NNS⟩, and ⟨NN NNS SENT⟩. Examining general counts over all unigram noun-related POS tags, i.e. ⟨NN⟩, ⟨NNS⟩, ⟨NP⟩, ⟨NPS⟩, returns somewhat inconclusive results. Both authors show a decrease for ⟨NNS⟩, and Twain also for ⟨NPS⟩. In contrast, there is a slight increase in pronouns in Twain's data. Overall, this might indicate a shift in how noun phrases are commonly constructed, i.e. simpler or more pronoun-based. Merely summing the tags does not

Figure 11: Prominent features common to Twain, James, and the reference corpus

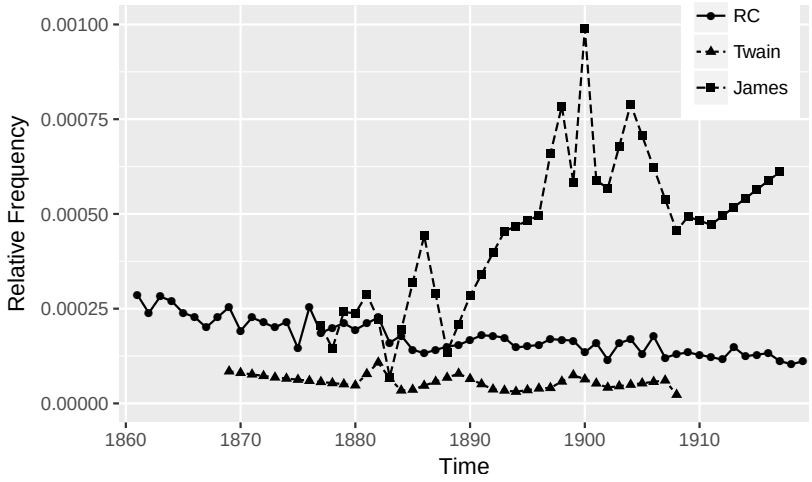


(a) Frequency of ⟨CC EX⟩ for James, Twain, and the RC

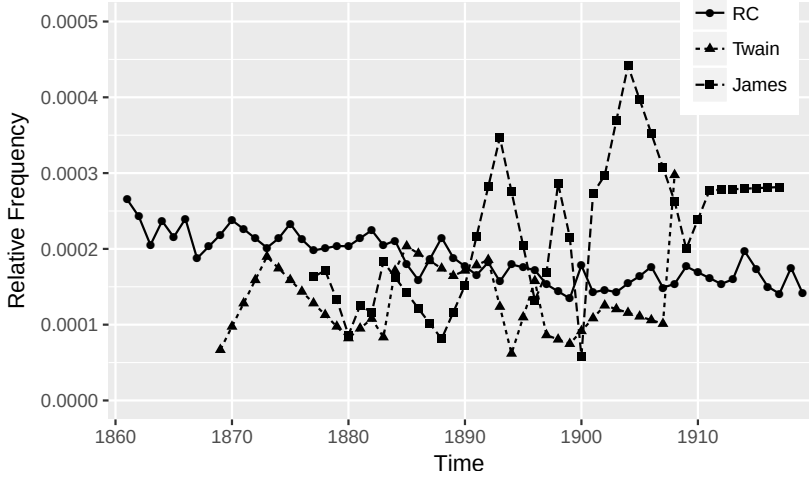


(b) Frequency of ⟨, by⟩ for James, Twain, and the RC

adequately describe how many noun phrases there are, nor how they are composed. Nor would simply looking at a rise or decrease in determiners suffice to ascertain how the above items are distributed. This result can only provide pointers for interesting aspects to consider in future work, which would require actually looking at the number of noun phrases overall and investigating whether the way they are composed changes over time.



(a) Frequency of ⟨WDT ,⟩ for James, Twain, and the RC



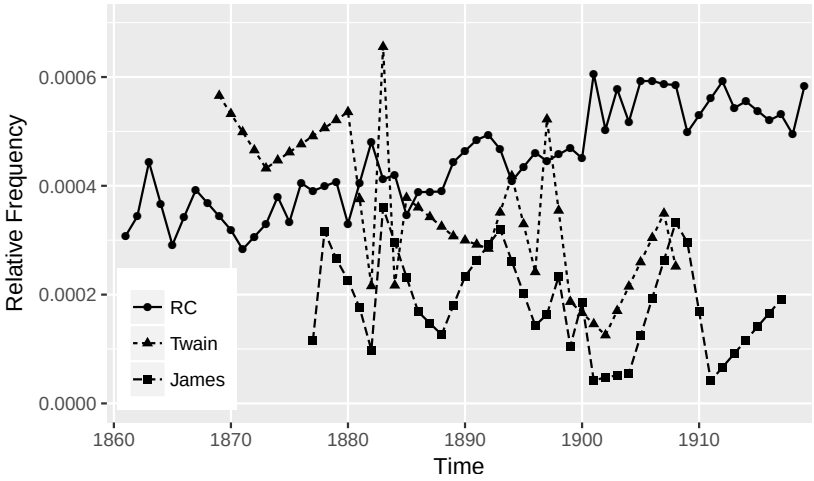
(b) Frequency of ⟨MD ,⟩ for James, Twain, and the RC

Figure 12: Prominent features common to Twain, James, and the reference corpus

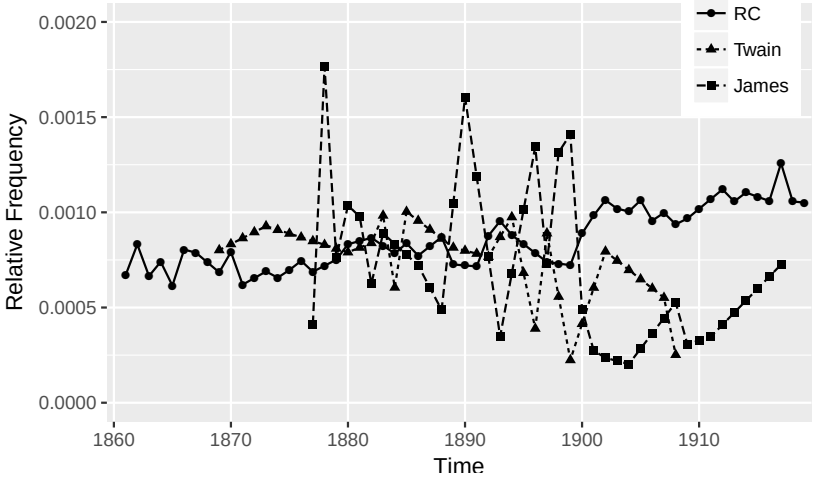
6.2.3 Stylistic differences between authors

In this final part, we consider some stylistic differences between the two authors. Although the graphs in Figure 11 have already been discussed as part of the comparison between reference corpus and author-specific models, these features are also interesting to analyse in terms of what this difference in usage implies about differences in authorial style. Three of these features (⟨WDT ,⟩, ⟨MD ,⟩, ⟨, by⟩) are certainly

Figure 13: Decrease in two noun phrase types for the two-author data set and increase for the reference set



(a) Frequency of ⟨DT NN NNS⟩ for James, Twain, and the RC



(b) Frequency of ⟨NN NN SENT⟩ for James, Twain, and the RC

more important for James, as Twain's usage mostly lies below that in the reference corpus. Examining some of the lexical realizations for these features for James and Twain shows clear differences in usage. James seems to use these features to build longer and more complicated sentences, increasing the cognitive workload on the part of the reader, which probably contributed to James' later style being considered somewhat 'obscure' and 'over-planned' (Beach 1918); an example of ⟨WDT ,⟩ is shown in (4).

(4) *It sounds, no doubt, too penetrating, but it was by no means all through Sir Claude's betrayals that Maisie was able to piece together the beauty of the special influence through which, for such stretches of time, he had refined upon propriety by keeping so far as as possible his sentimental interests distinct.*

There are a few instances of simpler constructions, not introducing a proper sub-clause, such as 'of which, however, she had', but these examples appear to be less numerous overall. While Twain's texts do contain these types of constructions, they appear more sparingly and also take a different, less convoluted shape, an example of which is shown in (5).

(5) *There is only a plausible resemblance, which, while it is apt enough to mislead the ignorant, cannot deceive parties who have contemplated both tribes.*

(6) *Then it is, in the final situation, that we get, by a backward reference or action, the real logic and process of the ambassador's view of how it has seemed best to take the thing, and what it...*

(7) *Without suspecting it, Dr. Peake, by entering the place, had reminded me of the talk of three years before.*

Examples of the syntactic word bigram ⟨ , by⟩ are shown in (6) and (7) for James and Twain, respectively.

7 DISCUSSION

This work has presented various experiments and analyses aimed at discovering salient features of general and individual language change. To identify these features, we used linear regression models, retaining only constant features for the reference corpus models, and shared constant features for the two-author models. Selecting only constant features serves to focus the analysis on the features the authors remained true to over their creative life span. Features used in a non-constant fashion would be interesting to analyse to complement the current results. We chose to only use linear models, for our experiments here, to limit the quantity of results. Other types of models should be studied in future work. As we chose to consider different feature types and n-gram sizes, there were many results and interpre-

tations to consider, and unfortunately we could not do justice to them all. The interpretations that we have provided are subjective, yet anchored in the critical literature that we have explored to date. We hope that other researchers will identify other natural categories within the features marked as salient by our methods, which may support competing interpretations. Our task in this work is not to propose definitive interpretations, but to provide methods to highlight features that undergo interesting development during writers' careers and to suggest that these interpretations may be anchored in critical responses to the career.

In terms of general differences from the reference corpus, there seems to be an interesting shift for both authors towards the use of simpler noun phrase constructions. We could not clearly identify all the particulars as part of this work. It would probably not suffice to simply analyse the composition of noun phrases, as genre and authorship could play a factor in this as well. One would therefore need to consider other contemporaneous authors to investigate the spread of this shift. In terms of more specific stylistic differences, we were able to find some common trends in both James and Twain, not found in the reference language, such as a decrease in the use of body references and a very marked difference for plural cases. This could suggest that James and Twain focus much more on the individual than was common for their time, but also that this particularity decreased over time. Existential constructions seemed to have generally gained more popularity over time in all three sets, with this being particularly pronounced in James and Twain.

Our analysis of Henry James and Mark Twain with a focus on stylistic changes has highlighted a number of differences between them, as for instance their use of female pronouns. James seems to have been highly progressive in his focus on the female perspective. This view is also supported by Baym (1981), who believes that James posed a continual challenge to the masculinist bias of American critical theory. An interesting aspect to consider as part of this investigation would be to compare James' style to a British reference corpus, given that he spent the latter part of his life in Europe.

In terms of syntactic style, there are a number of differences, one of which being that James seems to compose much more intricate sentences than Twain, especially towards the end of his life, as has

already been identified by literary scholars. In general, Twain's language is more pessimistic, questioning, and contains many more religious references than James' texts. From a more topic-based point of view, one might also consider frequent themes discussed as part of their works and possible changes in them over time. Overall, what one might say about Twain and James is that although they appear to often use the same tools, they apply them very differently. Regarding general differences from the reference corpus, it is probable that James and Twain did not really conform to the language of their time, although this would need to be verified by looking at the works of authors with comparable lifespans.

8 CONCLUSION

This work considered salient features of language change in the works of two prominent American authors, Henry James and Mark Twain, as well as in a reference corpus. We were able to identify a number of interesting changes in both lexical and syntactic features, suggesting other possible leads to explore. As style is a very general concept encompassing a multitude of possible dimensions, we were only able to 'scratch the surface', and more experiments should follow, to continue this work. The earlier part of this paper outlines only one method of discovery for salient features, but others should be considered and investigated. This work highlights the importance of using a reference corpus to verify that any change perceived in an author's style is indeed only to be found in the work of that author.

ACKNOWLEDGEMENT

We would like to thank our anonymous reviewers for their helpful suggestions on how to improve the earlier version of this paper. Further, we would also like to thank Carmela Chateau Smith, whose thorough work has greatly improved this paper's readability and consistency. This research is supported by Science Foundation Ireland (SFI) through the CNGL Programme (Grants 12/CE/I2267 and 13/RC/2106) in the ADAPT Centre (`www.adaptcentre.ie`).

REFERENCES

Alex AYRES (2010), *The Wit and Wisdom of Mark Twain*, Harper Collins.

Nina BAYM (1981), Melodramas of Beset Manhood: How Theories of American Fiction Exclude Women Authors, *American Quarterly*, 33(2):123–139, ISSN 00030678, 10806490, http://www.jstor.org/stable/2712312.

Joseph Warren BEACH (1918), *The Method of Henry James*, Yale University Press.

Walter BLAIR (1963), Reviewed Work: Twain and the Image of History by Roger B. Salomon, *American Literature*, 34(4):578–580, http://www.jstor.org/stable/2923090.

Van Wyck BROOKS (1920), The Ordeal of Mark Twain, *New York: Dutton*.

Henry Seidel CANBY (1951), *Turn West, Turn East: Mark Twain and Henry James*, Biblo & Tannen Publishers.

Walter DAELEMANS (2013), Explanation in Computational Stylometry, in *Computational Linguistics and Intelligent Text Processing*, pp. 451–462, Springer.

Mark DAVIES (2012), The 400 Million Word Corpus of Historical American English (1810–2009), in *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical linguistics (ICeHl 16), Pécs, 23–27 August 2010*, pp. 231–61.

Maciej EDER, Mike KESTEMONT, and Jan RYBICKI (2013), Stylometry with R: A Suite of Tools, in *Digital Humanities 2013: Conference Abstracts*, pp. 487–89, University of Nebraska–Lincoln, Lincoln, NE, http://dh2013.unl.edu/abstracts/.

Jerome FRIEDMAN, Trevor HASTIE, and Robert TIBSHIRANI (2001), *The Elements of Statistical Learning*, volume 1, Springer Series in Statistics Springer, Berlin.

Jerome FRIEDMAN, Trevor HASTIE, and Robert TIBSHIRANI (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1):1–22, http://www.jstatsoft.org/v33/i01/.

David L HOOVER (2007), Corpus Stylistics, Stylometry, and the Styles of Henry James, *Style*, 41(2):174–203.

Gareth JAMES, Daniela WITTEN, Trevor HASTIE, and Robert TIBSHIRANI (2013), *An Introduction to Statistical Learning*, volume 112, Springer.

Henry JAMES (1884), *The Art of Fiction*, Longmans, Green and Company.

Timothy P. JURKA, Loren COLLINGWOOD, Amber E. BOYDSTUN, Emiliano GROSSMAN, and Wouter VAN ATTEVELDT (2012), *RTextTools: Automatic Text Classification via Supervised Learning*, http://CRAN.R-project.org/package=RTextTools, R package version 1.3.9.

Michael J. KANE, John EMERSON, and Stephen WESTON (2013), Scalable Strategies for Computing with Massive Data, *Journal of Statistical Software*, 55(14):1–19, http://www.jstatsoft.org/v55/i14/.

Carmen KLAUSSNER and Carl VOGEL (2015), Stylochronometry: Timeline Prediction in Stylometric Analysis, in Max BRAMER and Miltos PETRIDIS, editors, *Research and Development in Intelligent Systems XXXII*, pp. 91–106, Springer International Publishing, Cham.

Moshe KOPPEL, Jonathan SCHLER, and Shlomo ARGAMON (2011), Authorship Attribution in the Wild, *Language Resource Evaluation*, 45(1):83–94, doi:10.1007/s10579-009-9111-2, http://dx.doi.org/10.1007/s10579-009-9111-2.

Moshe KOPPEL, Jonathan SCHLER, and Elisheva BONCHEK-DOKOW (2007), Measuring Differentiability: Unmasking Pseudonymous Authors, *Journal of Machine Learning Resources*, 8:1261–1276, ISSN 1532-4435, http://dl.acm.org/citation.cfm?id=1314498.1314541.

Max KUHN (2014), *Caret: Classification and Regression Training*, http://CRAN.R-project.org/package=caret, with contributions from: Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer and the R Core Team, R package version 6.0-30.

Spyros MAKRIDAKIS, Steven C WHEELWRIGHT, and Rob J HYNDMAN (2008), *Forecasting Methods and Applications*, John Wiley & Sons.

Meik MICHALKE (2014), *koRpus: An R Package for Text Analysis*, http://reaktanz.de/?c=hacking&s=koRpus, (Version 0.05-4).

James W PENNEBAKER and Lori D STONE (2003), Words of Wisdom: Language Use Over the Life Span, *Journal of Personality and Social Psychology*, 85(2):291–231.

REVOLUTION ANALYTICS and Steve WESTON (2014), *foreach: Foreach looping construct for R*, http://CRAN.R-project.org/package=foreach, R package version 1.4.2.

Paolo ROSSO, Francisco M. Rangel PARDO, Martin POTTHAST, Efstathios STAMATATOS, Michael TSCHUGGNALL, and Benno STEIN (2016), Overview of PAN'16 – New Challenges for Authorship Analysis: Cross-Genre Profiling, Clustering, Diarization, and Obfuscation, in *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5–8, 2016, Proceedings*, pp. 332–350.

Helmut SCHMID (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, in *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pp. 44–49, Manchester, UK.

Joseph A. SMITH and Colleen KELLY (2002), Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works, *Computers and the Humanities*, 36(4):411–430, http://www.jstor.org/stable/30204686.

Efstathios STAMATATOS (2012), On the Robustness of Authorship Attribution Based on Character N-gram Features, *Journal of Law & Policy*, 21:421–439.

THOMAS M. WALSH AND THOMAS D. ZLATIC (1981), Mark Twain and the Art of Memory, *American Literature*, 53(2):214–231, http://www.jstor.org/stable/2926100.

James D. WILLIAMS (1965), The Use of History in Mark Twain's 'A Connecticut Yankee', *PMLA*, 80(1):102–110, http://www.jstor.org/stable/461131.

Hui ZOU and Trevor HASTIE (2005), Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, http://www.jstor.org/stable/3647580.